Submitted to the Bernoulli
arXiv: math.PR/0000000

Posterior convergence rates in non-linear latent variable models

DEBDEEP PATI[*], ANIRBAN BHATTACHARYA[**] and DAVID DUNSON[†]

Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251, USA E-mail: [*]dp55@stat.duke.edu; [**]ab179@stat.duke.edu; [†]dunson@stat.duke.edu

Non-linear latent variable models have become increasingly popular in a variety of applications. However, there has been little study on theoretical properties of these models. In this article, we study rates of posterior contraction in univariate density estimation for a class of non-linear latent variable models where unobserved $\mathrm{U}(0,1)$ latent variables are related to the response variables via a random non-linear regression with an additive error. Our approach relies on characterizing the space of densities induced by the above model as kernel convolutions with a general class of continuous mixing measures. The literature on posterior rates of contraction in density estimation almost entirely focuses on finite or countably infinite mixture models. We develop approximation results for our class of continuous mixing measures. Using an appropriate Gaussian process prior on the unknown regression function, we obtain the optimal frequentist rate up to a logarithmic factor under standard regularity conditions on the true density.

AMS 2000 subject classifications: Primary 62G07, 62G20; secondary 60K35.
Keywords: Bayesian nonparametrics, Density estimation, Gaussian process, Maximum entropy moment-matching, One factor model, Rate of convergence.

1. Introduction

Kernel mixture models are known to be extremely flexible and have been extensively used for density estimation. Starting with a parametric kernel $\mathcal{K}(y,\theta)$, one can obtain a class of densities f_G as

$$f_G(y) = \int \mathcal{K}(y,\theta)dG(\theta), \tag{1.1}$$

where $G(\cdot)$ is a mixing distribution. In particular, by choosing G to be a discrete distribution with finitely many atoms $\theta_h, h = 1,\ldots,k$ having weights $\pi_h, h = 1,\ldots,k$ with $\sum_{h=1}^{k}\pi_h = 1$, one obtains the important class of finite mixture models. In a Bayesian framework, one can induce a prior distribution on the class of densities by assigning a prior to G, which amounts to specifying priors on k and $(\theta_h,\pi_h), h = 1,\ldots,k$ in case of finite mixture models. A Dirichlet process (Ferguson, 1973, 1974) is often used as a default prior on the class of mixing distributions due to its attractive theoretical properties and availability of efficient algorithms for posterior computation. Since realizations of a Dirichlet process are almost surely discrete (see Sethuraman (1994) for a constructive

definition), a Dirichlet process prior on G induces an infinite discrete mixture model for f_G. A well known drawback of finite mixture models is the sensitivity of the results to the choice of k, whereas updating k in a fully Bayesian formulation is computationally intensive. The infinite mixture representation avoids fixing a truncation level and sophisticated sampling algorithms such as Walker (2007) enable posterior sampling from the full posterior distribution.

Although finite and infinite discrete mixture models have been extensively used, there are reasons to look beyond these classes of models. A discrete prior on G partitions the n subjects into one or more clusters, with subjects in the same cluster sharing the same θ value. Although this property has been widely exploited for probabilistic clustering, one might want to avoid the clustering phenomenon in situations where the interest is purely in density estimation and one is not interested in interpreting the clusters or in inferring the cluster specific parameters. It is often the case that the clusters don't have any physical significance and subjects get inappropriately grouped together for all parameter values obscuring subtle differences. In such cases, the clustering is more of an artifact of the model and a continuum among the parameter values for the subjects seems more reasonable.

While Polya tree priors (Ferguson, 1974; Mauldin, Sudderth and Williams, 1992) can be directly used to induce priors on the space of absolutely continuous densities (Lavine, 1992), the resulting density estimates are found to be spiky in practice. Lenk (1988, 1991) proposed a logistic Gaussian process which bypasses the mixture formulation by directly modeling an unknown density on the unit interval as the exponent of a random function re-normalized, or equivalently modeling the log-density using a Gaussian process prior. The normalizing constant in the logistic Gaussian process models is analytically intractable and causes difficulties in posterior sampling. Refer to Tokdar (2007) for a faster implementation in density estimation with logistic Gaussian process priors.

Recently, Kundu and Dunson (2011) proposed an approach for univariate density estimation in which the response variables are modeled as unknown functions of uniformly distributed latent variables with an additive Gaussian error. The latent variable specification allows straightforward posterior computation via conjugate posterior updates. Since inverse c.d.f. transforms of uniform random variables can generate draws from any distribution, by choosing the prior on the error variance to assign positive mass to arbitrary neighborhoods of zero while placing a prior with large support on the space of functions mapping the latent variables to the observed variables (referred to as the *transfer function* from now on), their prior can approximate draws from any continuous distribution function arbitrarily closely. One can also conveniently center the non-parametric model on a parametric family by centering the prior on the transfer function on a parametric class of quantile (or inverse c.d.f.) functions $\{F_\theta^{-1} \ : \ \theta \in \Theta\}$. While such centering on parametric guesses can be achieved in Dirichlet process mixture models by appropriate choice of the base measure G_0, posterior computation becomes complicated unless the base measure is conjugate to the kernel $\mathcal{K}$.

There has been growing interest in studying asymptotic properties of Bayesian procedures assuming the data are sampled from a fixed unknown distribution. The posterior distribution is said to be strongly consistent if it concentrates almost surely in arbi-

trarily small L_1 neighborhoods of the true distribution with increasing sample size. Ghosal, Ghosh and Ramamoorthi (1999) provided general conditions in terms of L_1 metric entropy to ensure strong posterior consistency and verified those conditions for Dirichlet process location mixtures of normal kernels under certain regularity conditions. Tokdar (2006) extended their result to the location-scale mixture case while encompassing a significantly larger class of "true" densities. Ghosal, Ghosh and van der Vaart (2000) considered the rate of contraction of a posterior distribution to the true density, providing an upper bound on the rate at which one can let the neighborhood size decrease to zero. Ghosal and van der Vaart (2001) obtained rates of posterior contraction for the Dirichlet process mixture model when the true density is a location-scale mixture of normals with component specific standard deviations bounded between two positive numbers. Although a nearly parametric rate is obtained in this case, the above class of densities is restrictive since one needs the component specific standard deviations to be arbitrarily small for normal mixtures to be able to approximate any smooth density. Ghosal and van der Vaart (2007) developed a generalization of the basic rate theorem in Ghosal, Ghosh and van der Vaart (2000) and addressed a broader class of densities, namely, the class of twice continuously differentiable densities. Under some regularity conditions which include the requirement that the true density be compactly supported, they obtained the optimal minimax rate of $n^{-2/5}$ up to a logarithmic factor based on Dirichlet process mixture models. Kruijer, Rousseau and van der Vaart (2010) considered finite location-scale mixtures of exponential power distributions and obtained minimax rates of convergence up to a logarithmic factor for any β-Hölder density, implying rate adaptivity to any degree of smoothness of the true density.

In this article, we study rates of posterior contraction in univariate density estimation for a class of non-linear latent variable models (NL-LVM) similar to Kundu and Dunson (2011). The NL-LVM encompasses a large class of univariate densities and it is straightforward to extend the class for multivariate density estimation and density regression problems. In particular, the NL-LVM has elements in common to Gaussian process latent variable models (GP-LVM) routinely used in machine learning applications for high-dimensional data visualization and dimensionality reduction (Lawrence, 2004, 2005; Lawrence and Moore, 2007; Ferris, Fox and Lawrence, 2007). However, the literature on GP-LVM doesn't provide any discussion on the flexibility of their specification in terms of the induced density of the observations after marginalizing out the latent variables. Although Kundu and Dunson (2011) provide an intuitive argument for large support in the density space for the univariate case, a rigorous characterization of the prior support is missing. We provide an accurate characterization of the prior support in terms of kernel convolution with a class of continuous mixing measures. We provide conditions for the mixing measure to admit a density with respect to Lebesgue measure and show that the prior support of the NL-LVM is at least as large as that of DP mixture models. We then develop approximation results for the above class of continuous mixing measures and subsequently derive posterior contraction rates assuming standard smoothness assumptions on the true density. Assuming the true density to be twice continuously differentiable, the best obtainable rate is found to be the minimax rate of $n^{-2/5}$ up to a logarithmic factor. Further, if the prior on the transfer function is centered on a paramet-

ric family which happens to contain the true density, then one gets a faster convergence rate which can be arbitrarily close to the parametric rate of $n^{-1/2}$ up to a logarithmic factor. Also, analogous to the Dirichlet process mixture models, when the true density is a Gaussian convolution with a finite mixture of truncated Gaussians, one can also attain a near parametric convergence rate.

The main contributions of this article are as follows. (i) The characterization of our model using convolutions implies that one can approximate any continuous density by choosing the transfer function to be the quantile function of the true density and letting the error variance to decrease to zero. When the true density is not compactly supported, the corresponding quantile function is unbounded with discontinuities at 0 and 1 and it is not immediate whether a prior for the transfer function supported on $C[0,1]$ (a default choice being a Gaussian process prior) results in the optimal rate. To address this issue, we define a sequence of $C[0,1]$ functions that converge pointwise to the true quantile function and derive concentration bounds for the prior around this sequence. (ii) The traditional approach of approximating the Gaussian convolution of a compactly supported density by discrete normal mixtures isn't well-suited for our purpose since the quantile function of the mixing distribution is a step function which doesn't belong to the sup-norm support of any smooth stochastic process. We develop a technique based on maximum entropy moment matching (Mead and Papanicolaou, 1984) for approximating a compactly supported density by an infinitely smooth density. Although the above developments are crucially used for our treatment of the non-compact case, we believe these results will be of independent interest.

The rest of the article is organized as follows. We introduce relevant notations and terminologies in Section 2. To make the article self-contained, we also provide a brief background on Gaussian process priors. In Section 3, we formulate our assumptions on the true density f_0 and in the following section, we describe the NL-LVM model and relate it to convolutions. We state our main theorem on convergence rates for the compact case in Section 5 and the non-compact case in Section 6. We discuss some special cases in Section 7. Section 8 discusses some implications of our results and outlines possible future directions.

2. Notations

Throughout the article, $Y_1, \ldots, Y_n, \ldots$ are independent and identically distributed with density $f_0 \in \mathcal{F}$, the set of all densities on $\mathbb{R}$ absolutely continuous with respect to the Lebesgue measure λ. The supremum and L_1-norm are denoted by $\|\cdot\|_\infty$ and $\|\cdot\|_1$, respectively. We let $\|\cdot\|_{p,\nu}$ denote the norm of $L_p(\nu)$, the space of measurable functions with ν-integrable pth absolute power. For two density functions $f, g \in \mathcal{F}$, let h denote the Hellinger distance defined as $h^2(f,g) = \left\|\sqrt{f} - \sqrt{g}\right\|_{2,\lambda} = \int (f^{1/2} - g^{1/2})^2 d\lambda$, $K(f,g)$ the Kullback-Leibler divergence given by $K(f,g) = \int \log(f/g) f d\lambda$ and $V(f,g) = \int \log(f/g)^2 f d\lambda$. The notation $C[0,1]$ is used for the space of continuous functions $f : [0,1] \to \mathbb{R}$ endowed with the supremum norm. For $\beta > 0$, we let $C^\beta[0,1]$ denote the Hölder space of order β, consisting of the functions $f \in C[0,1]$ that have $\lfloor \beta \rfloor$ continuous

derivatives with the $\lfloor\beta\rfloor$th derivative $f^{\lfloor\beta\rfloor}$ being Lipschitz continuous of order $\beta - \lfloor\beta\rfloor$. The ϵ-covering number $N(\epsilon, S, d)$ of a semi-metric space S relative to the semi-metric d is the minimal number of balls of radius ϵ needed to cover S. The logarithm of the covering number is referred to as the entropy. By near-optimal rate of convergence we mean optimal rate of convergence slowed down by a logarithmic factor.

We write "$\precsim$" for inequality up to a constant multiple. Let $\phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$ denote the standard normal density, and let $\phi_\sigma(x) = (1/\sigma)\phi(x/\sigma)$. Let an asterisk denote a convolution e.g., $(\phi_\sigma * f)(y) = \int \phi_\sigma(y-x)f(x)dx$. The support of a density f is denoted by $\mathrm{supp}(f)$.

We briefly recall the definition of the RKHS of a Gaussian process prior; a detailed review can be found in van der Vaart and van Zanten (2008b). A Borel measurable random element W with values in a separable Banach space $(\mathbb{B}, \|\cdot\|)$ (e.g., $C[0,1]$) is called Gaussian if the random variable b^*W is normally distributed for any element $b^* \in \mathbb{B}^*$, the dual space of $\mathbb{B}$. The reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ attached to a zero-mean Gaussian process W is defined as the completion of the linear space of functions $t \mapsto EW(t)H$ relative to the inner product

$$\langle \mathrm{E}W(\cdot)H_1; \mathrm{E}W(\cdot)H_2\rangle_{\mathbb{H}} = \mathrm{E}H_1H_2,$$

where H, H_1 and H_2 are finite linear combinations of the form $\sum_i a_i W(s_i)$ with $a_i \in \mathbb{R}$ and s_i in the index set of W.

3. Assumptions on the true density

It has been widely recognized that one needs certain smoothness assumptions and tail conditions on the true density f_0 to derive posterior convergence rates at f_0. We need the following assumptions in our case,

Assumption 3.1. *f_0 is twice continuously differentiable with $\int (f_0''/f_0)^2 f_0 d\lambda < \infty$ and $\int (f_0'/f_0)^4 f_0 d\lambda < \infty$.*

***Remark* 3.1.** Letting $f_0(y) = C\exp\{-w_0(y)\}$ on $\mathrm{supp}(f_0)$ so that $w_0 = \log C - \log f_0(y)$, we can restate Assumption 3.1 as w_0 being twice continuously differentiable and

$$\int_{-\infty}^{\infty} \{w_0'(y)\}^4 \exp\{-w_0(y)\} < \infty, \int_{-\infty}^{\infty} \{w_0''(y)\}^2 \exp\{-w_0(y)\} < \infty. \tag{3.1}$$

Assumption 3.2. *f_0 is bounded, nondecreasing on $(-\infty, a]$, bounded away from 0 on $[a,b]$ and non-increasing on $[b,\infty)$ for some $a \le b$.*

Assumption 3.1 is the same as Assumption 1.2 of Ghosal and van der Vaart (2007) and ensures that $h(f_0, f_0 * \phi_\sigma) = O(\sigma^2)$ as $\sigma \to 0$; see Lemma 4 of Ghosal and van der Vaart (2007) for a proof. Assumption 3.2 is the same as the assumption in Lemma 6 of the same

paper. This is sufficient to guarantee that for every $\delta > 0$, there exists a constant $C > 0$ such that $f_0 * \phi_\sigma \geq C f_0$ for every $\sigma < \delta$. While Assumption 3.1 only allows sufficiently smooth densities, Assumption 3.2 is only a mild requirement in the sense that most reasonable densities arising in practice should satisfy it. Moreover, if f_0 is nondecreasing on $(-\infty, a]$ and nonincreasing on $[b, \infty]$ for some $a \leq b$, f_0 is automatically bounded and bounded away from zero on $[a, b]$ provided it is continuous and no-where zero on $[a, b]$.

4. The NL-LVM model

Consider the nonlinear latent variable model,

$$y_i = \mu(\eta_i) + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2), \ (i = 1, \ldots, n) \tag{4.1}$$

$$\mu \sim \Pi_\mu, \quad \sigma \sim \Pi_\sigma, \quad \eta_i \sim \mathrm{U}(0, 1), \tag{4.2}$$

where η_i's are subject specific latent variables, $\mu \in C[0, 1]$ is a *transfer function* relating the latent variables to the observed variables and ϵ_i is an idiosyncratic error specific to subject i. The density of y conditional on the transfer function μ and scale σ is obtained on marginalizing out the latent variable as

$$f(y; \mu, \sigma) \stackrel{\text{def}}{=} f_{\mu,\sigma}(y) = \int_0^1 \phi_\sigma(y - \mu(x))dx. \tag{4.3}$$

Define a map $g : C[0, 1] \times [0, \infty) \to \mathcal{F}$ with $g(\mu, \sigma) = f_{\mu,\sigma}$. One can induce a prior Π on $\mathcal{F}$ via the mapping g by placing independent priors Π_μ and Π_σ on $C[0, 1]$ and $[0, \infty)$ respectively, with $\Pi = (\Pi_\mu \otimes \Pi_\sigma) \circ g^{-1}$. Kundu and Dunson (2011) assumed a Gaussian process prior with squared exponential covariance kernel on μ and an inverse-gamma prior on σ^2.

It is not immediately clear whether the class of densities $f_{\mu,\sigma}$ in the range of g encompass a large subset of the density space. We provide an intuition that relates the above class with convolutions and is crucially used later on. Let f_0 be a continuous density with cumulative distribution function $F_0(t) = \int_{-\infty}^t f_0(x)dx$. Assume f_0 to be non-zero almost everywhere within its support, so that $F_0 : \mathrm{supp}(f_0) \to [0, 1]$ is strictly monotone and hence has an inverse $F_0^{-1} : [0, 1] \to \mathrm{supp}(f_0)$ satisfying $F_0\{F_0^{-1}(t)\} = t$ for all $t \in \mathrm{supp}(f_0)$. If $\mathrm{supp}(f_0) = \mathbb{R}$, then the domain of F_0^{-1} is the open interval $(0, 1)$ instead of $[0, 1]$.

Letting $\mu_0(x) = F_0^{-1}(x)$, one obtains

$$f_{\mu_0,\sigma}(y) = \int_0^1 \phi_\sigma(y - F_0^{-1}(x))dx = \int_{-\infty}^{\infty} \phi_\sigma(y - t) f_0(t)dt, \tag{4.4}$$

where the second equality follows from the change of variable theorem. Thus, $f_{\mu_0,\sigma}(y) = \phi_\sigma * f_0$, i.e., $f_{\mu_0,\sigma}$ is the convolution of f_0 with a normal density having mean 0 and standard deviation σ. It is well known that the convolution $\phi_\sigma * f_0$ can approximate f_0 arbitrary closely as the bandwidth $\sigma \to 0$. More precisely, for $f_0 \in L^p(\lambda)$ for any $p \geq 1$,

$\|\phi_\sigma * f_0 - f_0\|_{p,\lambda} \to 0$ as $\sigma \to 0$. Furthermore, a stronger result $\|\phi_\sigma * f_0 - f_0\|_\infty = O(\sigma^2)$ holds if f_0 is compactly supported. A similar result holds for the Hellinger metric, with the precise approximation error under Assumption 3.1 given by $h(\phi_\sigma * f_0, f_0) = O(\sigma^2)$ as $\sigma \to 0$.

Suppose the prior Π_μ on μ has full sup-norm support on $C[0,1]$ so that $\Pr(\|\mu - \mu^*\|_\infty < \epsilon) > 0$ for any $\epsilon > 0$ and $\mu^* \in C[0,1]$, and the prior Π_σ on σ has full support on $[0,\infty)$. If f_0 is compactly supported so that the quantile function $\mu_0 \in C[0,1]$, then it can be shown that under mild conditions, the induced prior Π assigns positive mass to arbitrarily small L_1 neighborhoods of any density f_0. When f_0 has full support on $\mathbb{R}$, the quantile function μ_0 is unbounded near 0 and 1, so that $\|\mu_0\|_\infty = \infty$. However, $\int_0^1 |\mu_0(t)|\, dt = \int_{\mathbb{R}} |x|\, f_0(x)dx$, which implies that μ_0 can be identified as an element of $L_1[0,1]$ if f_0 has finite first moment. Since $C[0,1]$ is dense in $L_1[0,1]$, the previous conclusion regarding L_1 support can be shown to hold in the non-compact case too. We summarize the above discussion in the following theorem, with a proof provided in the appendix.

Theorem 4.1. *If Π_μ has full sup-norm support on $C[0,1]$ and Π_σ has full support on $[0,\infty)$, then the L_1 support of the induced prior Π on $\mathcal{F}$ contains all densities f_0 which have a finite first moment and are non-zero almost everywhere on their support.*

***Remark* 4.1.** The conditions of Theorem 4.1 are satisfied for a wide range of Gaussian process priors on μ (for example, a GP with a squared exponential or Matérn covariance kernel).

Let $\tilde{\lambda}$ denote the Lebesgue measure on $[0,1]$, or equivalently, the U$[0,1]$ distribution. For any measurable function $\mu : [0,1] \to \mathbb{R}$, let ν_μ denote the induced measure on $(\mathbb{R}, \mathcal{B})$, with $\mathcal{B}$ denoting the Borel sigma-field on $\mathbb{R}$. Then, for any Borel measurable set B, $\nu_\mu(B) = \tilde{\lambda}(\mu^{-1}(B))$, where $\mu^{-1}(B) = \{x \in [0,1] \ : \ \mu(x) \in B\}$. By the change of variable theorem for induced measures,

$$\int_0^1 \phi_\sigma(y - \mu(x))dx = \int \phi_\sigma(y - t)d\nu_\mu(t), \tag{4.5}$$

so that $f_{\mu,\sigma}$ can be expressed as a kernel mixture form as in (1.1) with mixing distribution ν_μ. It turns out that this mechanism of creating random distributions is very general. Depending on the choice of μ, one can create a large variety of mixing distributions based on this specification. For example, if μ is a strictly monotone function, then ν_μ is absolutely continuous with respect to the Lebesgue measure, while choosing μ to be a step function, one obtains a discrete mixing distribution. However, it is easier to place a prior on μ supported on the space of continuous functions $C[0,1]$ without further shape restrictions and Theorem 4.1 assures us that this specification leads to large L_1 support on the space of densities.

5. The compact case

We first consider the case where f_0 is compactly supported, i.e., there exist $-\infty < a_0 < b_0 < \infty$ such that $\int_{a_0}^{b_0} f_0(x) = 1$. In that case, the quantile function $F_0^{-1} : [0,1] \to [a_0, b_0]$ is a continuous monotone function inheriting the smoothness of f_0. Denote the quantile function by μ_0. Assumption 3.1 ensures that the compactly supported density decays smoothly at the boundaries. Under Assumption 3.1 and the fundamental theorem of calculus, $\mu_0 : [0,1] \to [a_0, b_0]$ is thrice continuously differentiable implying $\mu_0 \in C^3[0,1]$.

5.1. Prior specification

We now mention our choices for the prior distributions Π_μ and Π_σ.

Assumption 5.1. *We assume μ follows a centered Gaussian process denoted by $GP(0, c)$, with a squared exponential covariance kernel $c(\cdot, \cdot; A)$ and a Gamma prior for the inverse-bandwidth A. Thus $c(t, s; A) = e^{-A(t-s)^2}, t, s \in [0,1], A \sim Ga(p, q)$.*

Assumption 5.2. *We assume $\sigma \sim IG(a_\sigma, b_\sigma)$.*

Note that contrary to the usual conjugate choice of an inverse-Gamma prior for σ^2, we have assumed an inverse-Gamma prior for σ. This enables one to have slightly more prior mass near zero compared to an inverse-Gamma prior for σ^2, leading to the optimal rate of posterior convergence. Refer also to Kruijer, Rousseau and van der Vaart (2010) for a similar prior choice for the bandwidth of the kernel in discrete location-scale mixture priors for densities.

5.2. Posterior convergence rate for the compact case

We state below the main theorem of posterior convergence rates.

Theorem 5.1. *If f_0 satisfies Assumption 3.1 and the priors Π_μ and Π_σ are as in Assumptions 5.1 and 5.2 respectively, the best obtainable rate of posterior convergence relative to h is*

$$\epsilon_n = n^{-\frac{2}{5}} \log n. \tag{5.1}$$

The proof of Theorem 5.1 is based on Ghosal and van der Vaart (2007) and van der Vaart and van Zanten (2007, 2008b, 2009). Unlike the treatment in discrete mixture models (Ghosal and van der Vaart, 2007) where a compactly supported density is approximated with a discrete mixture of normals, the main trick here is to approximate the true density f_0 by the convolution $\phi_\sigma * f_0$ and allow the prior on the transfer function to appropriately concentrate around the true quantile function $\mu_0 \in C[0,1]$.

To guarantee that the above scheme leads to the optimal rate of convergence, we first derive sharp bounds for the Hellinger distance between f_{μ_1,σ_1} and f_{μ_2,σ_2} for $\mu_1, \mu_2 \in C[0,1]$ and $\sigma_1, \sigma_2 > 0$. We summarize the result in the following Lemma 5.1.

Lemma 5.1. *For $\mu_1, \mu_2 \in C[0,1]$ and $\sigma_1, \sigma_2 > 0$,*

$$h^2(f_{\mu_1,\sigma_1}, f_{\mu_2,\sigma_2}) \leq 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} \exp\left\{ - \frac{\|\mu_1-\mu_2\|_\infty^2}{4(\sigma_1^2+\sigma_2^2)} \right\}. \tag{5.2}$$

Proof. Note that by Hölder's inequality,

$$f_{\mu_1,\sigma_1}(y) f_{\mu_2,\sigma_2}(y) \geq \left\{ \int_0^1 \sqrt{\phi_{\sigma_1}(y-\mu_1(x))}\sqrt{\phi_{\sigma_2}(y-\mu_2(x))}dx \right\}^2.$$

Hence,

$$\begin{aligned} h^2(f_{\mu_1,\sigma_1}, f_{\mu_2,\sigma_2}) &\leq \int \left[\int_0^1 \phi_{\sigma_1}(y-\mu_1(x))dx + \int_0^1 \phi_{\sigma_2}(y-\mu_2(x))dx \right. \\ &- \left. 2\int_0^1 \sqrt{\phi_{\sigma_1}(y-\mu_1(x))}\sqrt{\phi_{\sigma_2}(y-\mu_2(x))}dx \right] dy. \end{aligned}$$

By changing the order of integration (applying Fubini's theorem since the function within the integral is jointly integrable) we get,

$$\begin{aligned} h^2(f_{\mu_1,\sigma_1}, f_{\mu_2,\sigma_2}) &\leq \int_0^1 h^2(f_{\mu_1(x),\sigma_1}, f_{\mu_2(x),\sigma_2})dx \\ &= \int_0^1 \left[1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} \exp\left\{ - \frac{(\mu_1(x)-\mu_2(x))^2}{4(\sigma_1^2+\sigma_2^2)} \right\} \right] dx \\ &\leq 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}} \exp\left\{ - \frac{\|\mu_1-\mu_2\|_\infty^2}{4(\sigma_1^2+\sigma_2^2)} \right\}. \end{aligned}$$

□

***Remark* 5.1.** When $\sigma_1 = \sigma_2 = \sigma$, $h^2(f_{\mu_1,\sigma}, f_{\mu_2,\sigma}) \leq 1 - \exp\left\{ \|\mu_1-\mu_2\|_\infty^2 / 8\sigma^2 \right\}$, which implies that $h^2(f_{\mu_1,\sigma}, f_{\mu_2,\sigma}) \precsim \|\mu_1-\mu_2\|_\infty^2 / \sigma^2$.

***Remark* 5.2.** Note that if we had used $h^2(f_{\mu_1,\sigma_1}, f_{\mu_2,\sigma_2}) \leq \|f_{\mu_1,\sigma_1} - f_{\mu_2,\sigma_2}\|_1$, we would have obtained the cruder bound

$$h^2(f_{\mu_1,\sigma_1}, f_{\mu_2,\sigma_2}) \leq C_1 \frac{\|\mu_1-\mu_2\|_\infty}{(\sigma_1 \wedge \sigma_2)} + C_2 \frac{|\sigma_2-\sigma_1|}{(\sigma_1 \wedge \sigma_2)},$$

which is linear in $\|\mu_1-\mu_2\|_\infty$ for some constant $C_1, C_2 > 0$. This bound is less sharp than what is obtained in Lemma 5.1 and does not suffice for obtaining the optimal rate of convergence.

To control the Kullback-Leibler distance between the true density f_0 and the model $f_{\mu,\sigma}$, we derive an upper bound for $\log \left\| \frac{f_0}{f_{\mu,\sigma}} \right\|_\infty$ in Lemma 5.2.

Lemma 5.2. *If f_0 satisfies Assumption 3.2,*

$$\log \left\| \frac{f_0}{f_{\mu,\sigma}} \right\|_\infty \leq C_6 + \frac{\|\mu - \mu_0\|_\infty^2}{\sigma^2} \tag{5.3}$$

for some constant $C_6 > 0$.

Proof. Note that

$$\begin{aligned} f_{\mu,\sigma}(y) &= \frac{1}{\sqrt{2\pi}\sigma} \int_0^1 \exp\left\{ -\frac{(y-\mu(x))^2}{2\sigma^2} \right\} dx \\ &\geq \frac{1}{\sqrt{2\pi}\sigma} \int_0^1 \exp\left\{ -\frac{(y-\mu(x))^2}{\sigma^2} \right\} dx \exp\left\{ -\frac{\|\mu-\mu_0\|_\infty^2}{\sigma^2} \right\} \\ &\geq C_4 \phi_{\sigma/\sqrt{2}} * f_0(y) \exp\left\{ -\frac{\|\mu-\mu_0\|_\infty^2}{\sigma^2} \right\} \\ &\geq C_5 f_0(y) \exp\left\{ -\frac{\|\mu-\mu_0\|_\infty^2}{\sigma^2} \right\}, \end{aligned}$$

where the last inequality follows from Lemma 6 of Ghosal and van der Vaart (2007) by Assumption 3.2. Hence $\log \left\| \frac{f_0}{f_{\mu,\sigma}} \right\|_\infty \leq C_6 + \frac{\|\mu-\mu_0\|_\infty^2}{\sigma^2}$ for some constant $C_6 > 0$. □

***Remark* 5.3.** Note that if f_0 is compact then Assumption 3.2 is automatically satisfied.

Proof of Theorem 5.1: Following Ghosal, Ghosh and van der Vaart (2000), we need to find sequences $\bar{\epsilon}_n, \tilde{\epsilon}_n \to 0$ with $n \min\{\bar{\epsilon}_n^2, \tilde{\epsilon}_n^2\} \to \infty$ such that there exist constants $C_1, C_2, C_3, C_4 > 0$ and sets $\mathcal{F}_n \subset \mathcal{F}$ so that,

$$\log N(\epsilon_n, \mathcal{F}_n, d) \leq C_1 n \bar{\epsilon}_n^2 \tag{5.4}$$

$$\Pi(\mathcal{F}_n^c) \leq C_3 \exp\{-n\tilde{\epsilon}_n^2 (C_2 + 4)\} \tag{5.5}$$

$$\Pi\left(f_{\mu,\sigma} : \int f_0 \log \frac{f_0}{f_{\mu,\sigma}} \leq \tilde{\epsilon}_n^2, \int f_0 \log \left(\frac{f_0}{f_{\mu,\sigma}} \right)^2 \leq \tilde{\epsilon}_n^2 \right) \geq C_4 \exp\{-C_2 n \tilde{\epsilon}_n^2\}. \tag{5.6}$$

Then we can conclude that for $\epsilon_n = \max\{\bar{\epsilon}_n, \tilde{\epsilon}_n\}$ and sufficiently large $M > 0$, the posterior probability

$$\Pi_n(f_{\mu,\sigma} : d(f_{\mu,\sigma}, f_0) > M\epsilon_n | Y_1, \ldots, Y_n) \to 0 \text{ a.s. } P_{f_0}.$$

Let $W = (W_t : t \in \mathbb{R})$ be a Gaussian process with squared exponential covariance kernel. The spectral measure m_w of W is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}$ with the Radon-Nikodym derivative given by

$$\frac{dm_w}{d\lambda}(x) = \frac{1}{2\pi^{1/2}} e^{-x^2/4}.$$

Define a scaled Gaussian process $W^a = (W_{at} : t \in [0,1])$, viewed as a map in $C[0,1]$. Let $\mathbb{H}^a$ denote the RKHS of W^a, with the corresponding norm $\|\cdot\|_{\mathbb{H}^a}$. The unit ball in the RKHS is denoted $\mathbb{H}_1^a$. We will consider the Gaussian process $\mu \sim W^A$ given A, with $A \sim \text{Gamma}(p,q)$.

We will first verify (5.6) along the lines of Ghosal and van der Vaart (2007). Note that

$$h^2(f_0, f_{\mu,\sigma}) \precsim h^2(f_0, f_{\mu_0,\sigma}) + h^2(f_{\mu_0,\sigma}, f_{\mu,\sigma}). \tag{5.7}$$

Since $f_{\mu_0,\sigma} = \phi_\sigma * f_0$, using Lemma 4 of Ghosal and van der Vaart (2007), one obtains under Assumptions 3.1 and 3.2,

$$h^2(f_0, f_{\mu_0,\sigma}) \precsim O(\sigma^4). \tag{5.8}$$

From Lemma 5.1 and the following remark, we obtain

$$h^2(f_{\mu_0,\sigma}, f_{\mu,\sigma}) \precsim \frac{\|\mu - \mu_0\|_\infty^2}{\sigma^2}. \tag{5.9}$$

From Lemma 8 of Ghosal and van der Vaart (2007), one has

$$\int f_0 \log\left(\frac{f_0}{f_{\mu,\sigma}}\right)^i \le h^2(f_0, f_{\mu,\sigma})\left(1 + \log\left\|\frac{f_0}{f_{\mu,\sigma}}\right\|_\infty\right)^i \tag{5.10}$$

for $i = 1, 2$.

From (5.7)-(5.10), for any $b \ge 1$ and $\tilde{\epsilon}_n^2 = \sigma_n^4$,

$$\{\sigma \in [\sigma_n, \sigma_n + \sigma_n^b], \|\mu - \mu_0\|_\infty \precsim \sigma_n^3\} \subset$$
$$\left\{\int f_0 \log\frac{f_0}{f_{\mu,\sigma}} \precsim \sigma_n^4, \int f_0 \log\left(\frac{f_0}{f_{\mu,\sigma}}\right)^2 \precsim \sigma_n^4\right\}.$$

Then (5.6) will be satisfied with $\tilde{\epsilon}_n = n^{-\frac{2}{5}}$ if

$$\text{P}\{\sigma \in [\sigma_n, 2\sigma_n], \|\mu - \mu_0\|_\infty \precsim \sigma_n^3\} \ge \exp\{-C_4 n^{\frac{1}{5}}\}$$

for some constant $C_4 > 0$.

Since $\mu_0 \in C^3[0,1]$, from Section 5.1 of van der Vaart and van Zanten (2009),

$$\text{P}(\|\mu - \mu_0\|_\infty \le 2\delta_n) \ge C_5 \exp\{-C_6(1/\delta_n)^{1/3}\}(C_7/\delta_n)^{p/3},$$

for $\delta_n \to 0$ and constants $C_5, C_6, C_7 > 0$. Letting $\delta_n = \sigma_n^3$, we obtain

$$\text{P}(\|\mu - \mu_0\|_\infty \le 2\delta_n) \ge \exp\{-C_8(1/\sigma_n)\},$$

for some constant $C_8 > 0$. Since $\sigma \sim IG(a_\sigma, b_\sigma)$, we have

$$\begin{aligned} P(\sigma \in [\sigma_n, 2\sigma_n]) &= \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} \int_{\sigma_n}^{2\sigma_n} x^{-(a_\sigma+1)} e^{-b_\sigma/x} dx \\ &\geq \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} \int_{\sigma_n}^{2\sigma_n} e^{-2b_\sigma/x} dx \\ &\geq \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} \sigma_n \exp\{-b_\sigma/\sigma_n\} \\ &\geq \exp\{-C_9/\sigma_n\}, \end{aligned}$$

for some constant $C_9 > 0$. Hence

$$\mathrm{P}\{\sigma \in [\sigma_n, 2\sigma_n], \|\mu - \mu_0\|_\infty \precsim \sigma_n^3\} \geq \exp\{-C_4 n^{\frac{1}{5}}\},$$

with $\sigma_n = n^{-\frac{1}{5}}$, $\tilde{\epsilon}_n = n^{-\frac{2}{5}}$ and for some $C_4 > 0$.

Next we construct a sequence of subsets $\mathcal{F}_n$ such that 5.4 and 5.5 are satisfied with $\bar{\epsilon}_n = n^{-\frac{2}{5}} \log^{t_2} n$ and $\tilde{\epsilon}_n$ for some global constant $t_2 > 0$.

Letting $\mathbb{B}_1$ denote the unit ball of $C[0,1]$ and given positive sequences M_n, r_n, ξ_n, define

$$B_n = \left(M_n \sqrt{\frac{r_n}{\xi_n}} \mathbb{H}_1^r + \bar{\delta}_n \mathbb{B}_1 \right) \cup \left(\cup_{a<\xi_n} (M_n \mathbb{H}_1^a) + \bar{\delta}_n \mathbb{B}_1 \right)$$

as in van der Vaart and van Zanten (2009), with $\bar{\delta}_n = \bar{\epsilon}_n l_n / K_1$, $K_1 = 2(2/\pi)^{1/2}$ and let

$$\mathcal{F}_n = \{f_{\mu,\sigma} : \mu \in B_n, l_n < \sigma < h_n\}.$$

First we need to calculate $N(\bar{\epsilon}_n, \mathcal{F}_n, \|\cdot\|_1)$. Observe that for $\sigma_2 > \sigma_1 > \frac{\sigma_2}{2}$,

$$\|f_{\mu_1,\sigma_1} - f_{\mu_2,\sigma_2}\|_1 \leq \left(\frac{2}{\pi}\right)^{1/2} \frac{\|\mu_1 - \mu_2\|_\infty}{\sigma_1} + \frac{3(\sigma_2 - \sigma_1)}{\sigma_1}.$$

Taking $\kappa_n = \min\{\frac{\bar{\epsilon}_n}{6}, 1\}$ and $\sigma_m^n = l_n(1+\kappa_n)^m, m \geq 0$, we obtain a partition of $[l_n, h_n]$ as $l_n = \sigma_0^n < \sigma_1^n < \cdots < \sigma_{m_n-1}^n < h_n \leq \sigma_{m_n}^n$ with

$$m_n = \left(\log \frac{h_n}{l_n} \right) \frac{1}{\log(1+\kappa_n)} + 1. \tag{5.11}$$

One can show that $\frac{3(\sigma_m^n - \sigma_{m-1}^n)}{\sigma_{m-1}^n} = 3\kappa_n \leq \bar{\epsilon}_n/2$. Let $\{\tilde{\mu}_k^n, k = 1, \ldots, N(\bar{\delta}_n, B_n, \|\cdot\|_\infty)\}$ be a $\bar{\delta}_n$-net of B_n. Now consider the set

$$\{(\tilde{\mu}_k^n, \sigma_m^n) : k = 1, \ldots, N(\bar{\delta}_n, B_n, \|\cdot\|_\infty), 0 \leq m \leq m_n\}. \tag{5.12}$$

Then for any $f = f_{\mu,\sigma} \in \mathcal{F}_n$, we can find $(\tilde{\mu}_k^n, \sigma_m^n)$ such that $\|\mu - \tilde{\mu}_k^n\|_\infty < \bar{\delta}_n$. In addition, if one has $\sigma \in (\sigma_{m-1}^n, \sigma_m^n]$, then

$$\left\| f_{\mu,\sigma} - f_{\mu_k^n, \sigma_m^n} \right\|_1 \leq \bar{\epsilon}_n.$$

Hence the set in (5.12) is an $\bar{\epsilon}_n$-net of $\mathcal{F}_n$ and its covering number is given by

$$m_n N(\bar{\delta}_n, B_n, \|\cdot\|_\infty).$$

From the proof of Theorem 3.1 in van der Vaart and van Zanten (2009), for $\xi_n = \bar{\delta}_n/(2\tau M_n)$ and for any M_n, r_n with

$$M_n^{3/2}\sqrt{2\tau r_n} > 2\bar{\delta}_n^{3/2},\ r_n > a_0,\ M_n\sqrt{\|m_w\|} > \bar{\delta}_n, \tag{5.13}$$

we obtain

$$\log N(3\bar{\delta}_n, B_n, \|\cdot\|_\infty) \leq K_2 r_n \left(\log\left(\frac{M_n^{3/2}\sqrt{2\tau r_n}}{\bar{\delta}_n^{3/2}} \right) \right)^2 + 2\log \frac{2M_n\|m_w\|}{\bar{\delta}_n} \tag{5.14}$$

where $\tau^2 = \int_{\mathbb{R}} x^2 dm_w(x)$ and $\|m_w\|$ is the total variation norm of the spectral measure m_w.

Again from the proof of Theorem 3.1 in van der Vaart and van Zanten (2009), for sufficiently small $\bar{\delta}_n$ and $r_n > 1$ and for $M_n^2 > 16K_3 r_n(\log(r_n/\bar{\delta}_n))^2$, we have

$$P(W^A \notin B_n) \leq K_4 r_n^{p-1}\exp\{-K_5 r_n\} + \exp\{-M_n^2/8\} \tag{5.15}$$

for constants $K_3, K_4, K_5 > 0$.

Next we calculate $P(\sigma \notin [l_n, h_n])$. Observe that

$$\begin{aligned} P(\sigma \notin [l_n, h_n]) &= P(\sigma^{-1} < h_n^{-1}) + P(\sigma^{-1} > l_n^{-1}) \\ &\leq \sum_{k=\alpha_\sigma}^{\infty} \frac{e^{-b_\sigma h_n^{-1}}(b_\sigma h_n^{-1})^k}{k!} + \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} \int_{l_n^{-1}}^{\infty} e^{-b_\sigma x/2} dx \\ &\leq e^{-a_\sigma \log(h_n)} + \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} e^{-b_\sigma l_n^{-1}/2}. \end{aligned} \tag{5.16}$$

Thus with $h_n = O(\exp\{n^{1/5}\}), l_n = O(n^{-1/5}), r_n = O(n^{1/5}), M_n = O(n^{1/10}\log n)$, (5.15) and (5.16) implies

$$\Pi(\mathcal{F}_n^c) = \exp\{-K_6 n^{1/5}\}$$

for some constant $K_6 > 0$ guaranteeing that (5.5) is satisfied with $\tilde{\epsilon_n} = n^{-1/5}$.

Also with $\bar{\epsilon}_n = n^{-2/5}(\log n)$, (5.13) is satisfied and it follows from (5.11) and (5.14) that

$$\log N(\bar{\epsilon}_n, \mathcal{F}_n, \|\cdot\|_1) \leq K_7 n^{1/5}(\log n)^2$$

for some constant $K_7 > 0$.

Hence $\max\{\bar{\epsilon}_n, \tilde{\epsilon}_n\} = n^{-2/5}\log n$.

6. The non-compact case

The analysis of the non-compact case poses greater technical difficulties compared to the compact case, especially in verifying condition (5.6). Recall that in the compact case, $K(f_0, f_{\mu,\sigma}) \precsim O(\sigma^4) + \|\mu - \mu_0\|_\infty^2 / \sigma^2$. However, if $\mathrm{supp}(f_0) = \mathbb{R}$, then the corresponding quantile function μ_0 has $\|\mu_0\|_\infty = \infty$. This prohibits us from bounding $\int f_0 \log(f_0/f_{\mu,\sigma})^i, i = 1, 2$ using Lemma 5.2, since no prior for μ supported on $C[0,1]$ can concentrate around arbitrarily small neighborhoods of the true quantile function μ_0 in sup-norm. Since the tail behavior of f_0 has a one-to-one correspondence with the behavior of μ_0 near the boundary, we make additional assumptions on the tails of f_0 similar to (C3) in Kruijer, Rousseau and van der Vaart (2010).

Assumption 6.1. *f_0 has exponential tails, i.e., there exist positive constants T, M, τ_1, τ_2 such that*

$$f_0(x) \leq M \exp(-\tau_1 |x|^{\tau_2}), \quad |x| \geq T. \tag{6.1}$$

***Remark* 6.1.** Remark 3.1 suggests that under Assumption 6.1, w_0 behaves like a polynomial near the tails and hence Assumption 3.1 is automatically satisfied as long as w_0 or equivalently f_0 is twice continuously differentiable.

To derive concentration inequalities for the prior on μ, it is convenient to work with a series prior for μ as follows:

Assumption 6.2. *For an orthonormal basis $\{\phi_j\}_{j=0}^\infty$ of $L_2[0,1]$, a sequence of scales $\lambda_j \downarrow 0$, a fixed domain-rescaling integer a, a global scaling factor $b > 0$ and a truncation level J, consider a prior distribution for μ given by an orthonormal series expansion*

$$W^J(t) = \sum_{j=0}^{J} \lambda_j Z_j b \phi_j(at),\ t \in [0,1].$$

In the sequel we will chose sequences for J, a and b given by $J_n = O(n^{1/5}), b_n = n^{-1/10}(\log n)^{1/2}$ and $a_n = n^\alpha$ for some $\alpha > 1$ to attain the optimal rate of convergence.

***Remark* 6.2.** Let $W_{2,q}[0,1]$ denote the Sobolev space of $L_2[0,1]$ functions f whose weak partial derivative of order q, $D^q f \in L_2[0,1]$. Also, for $C > 0$, denote

$$W_{2,q}^C[0,1] = \{f \in L_2[0,1] : \|f\|_{2,q}^2 = \|D^q f\|_2^2 \leq C\} \tag{6.2}$$

to be the set of functions in $W_{2,q}[0,1]$ norm-bounded by C. In the sequel, we shall assume that ϕ_j's are given by a cosine basis

$$\phi_0(t) = \frac{1}{\sqrt{2}} \tag{6.3}$$

$$\phi_j(t) = \cos(2\pi j t),\ j \geq 1 \tag{6.4}$$

$$\tag{6.5}$$

so that for $f(t) = \sum_{j=0}^{\infty} \theta_j b \phi_j(at)$, one has $\|f\|_{2,q}^2 = b^2 \sum_{j=1}^{\infty} \theta_j^2 (2\pi a j)^q$. The techniques used subsequently can be easily extended to other orthonormal bases.

We are now in a position to state the main theorem of posterior convergence rates for the non-compact case.

Theorem 6.1. *If f_0 is twice continuously differentiable, satisfies Assumptions 3.2 and 6.1, and the priors Π_μ and Π_σ are as in Assumptions 6.2 and 5.2 respectively, the best obtainable posterior rate of convergence relative to h is*

$$\epsilon_n = n^{-\frac{2}{5}} (\log n)^{t_0}, \tag{6.6}$$

for some global constant t_0.

The construction of the sieves $\mathcal{F}_n$ is similar to the compact case and we shall omit the details of calculating the entropy and complement probability of $\mathcal{F}_n$ as they are essentially similar to the proof of Theorem 5.1. Verifying the KL condition in 5.6 is the biggest hurdle in the non-compact case; we briefly outline the steps needed to bound the integrals within parenthesis in 5.6. The basic idea is to separate the integrals into an integral over a compact set and its complement. Inside the compact set, one can replace f_0 by a compact approximation $f_{0\sigma}$ and approximate the quantile function $\mu_{0\sigma}$ of $f_{0\sigma}$ by an infinitely smooth function on $[0,1]$ in an appropriate sense, which enables one to obtain the right concentration rate using a smooth prior on $C[0,1]$. The complement term can be handled by exploiting the exponential tails of f_0.

To elaborate, first define sets $E_\sigma = \{x : f_0(x) > \sigma^{H_1}\}$, $E'_\sigma = \{x : f_0(x) > \sigma^{H_2}\}$. Clearly, $E_\sigma \subset E'_\sigma$ if $H_2 > H_1$. Without loss of generality, one can assume $E'_\sigma = [d_\sigma, e_\sigma]$ by Assumption 3.2. Let $g_{0\sigma} = f_0 1_{E'_\sigma}$ denote the restriction of f_0 to the compact set E'_σ and let $f_{0\sigma}$ be $g_{0\sigma}$ normalized to make it a density supported on E'_σ. Further, let $\mu_{0\sigma} : [0,1] \to E'_\sigma$ denote the quantile function of $f_{0\sigma}$ and denote $f_{\mu_{0\sigma},\sigma} = \phi_\sigma * f_{0\sigma}$.

We now bound $\mathrm{V}(f_0, f_{\mu,\sigma}) = \int f_0 \log(f_0/f_{\mu,\sigma})^2$, the treatment of $\mathrm{KL}(f_0, f_{\mu,\sigma})$ follows similarly. To start with, observe that

$$\int f_0 \log\left(\frac{f_0}{f_{\mu,\sigma}}\right)^2 \precsim \int f_0 \log\left(\frac{f_0}{f_{\mu_0,\sigma}}\right)^2 + \int f_0 \log\left(\frac{f_{\mu_0,\sigma}}{f_{\mu,\sigma}}\right)^2. \tag{6.7}$$

Using $f_{\mu_0,\sigma} = \phi_\sigma * f_0$, it follows from Lemma 8 of Ghosal and van der Vaart (2007) that,

$$\int f_0 \log\left(\frac{f_0}{f_{\mu_0,\sigma}}\right)^2 \le h^2(f_0, \phi_\sigma * f_0)\left(1 + \log\left\|\frac{f_0}{\phi_\sigma * f_0}\right\|_\infty\right)^2.$$

Since $h^2(f_0, \phi_\sigma * f_0) \precsim O(\sigma^4)$ from 5.8 and $\phi_\sigma * f_0 \ge C f_0$ by Assumption 3.2, one has $\int f_0 \log(f_0/f_{\mu_0,\sigma})^2 \precsim O(\sigma^4)$. To handle the second term in 6.7, we break up the integral

into integrals over E_σ and E_σ^c and further decompose the first term to obtain,

$$\int f_0 \log\left(\frac{f_{\mu_0,\sigma}}{f_{\mu,\sigma}}\right)^2 \precsim$$
$$\left\{\int_{E_\sigma} f_0 \log\left(\frac{f_{\mu_0,\sigma}}{f_{\mu_{0\sigma},\sigma}}\right)^2 + \int_{E_\sigma} f_0 \log\left(\frac{f_{\mu_{0\sigma},\sigma}}{f_{m_\sigma,\sigma}}\right)^2 + \int_{E_\sigma} f_0 \log\left(\frac{f_{m_\sigma,\sigma}}{f_{\mu,\sigma}}\right)^2\right\} + \int_{E_\sigma^c} f_0 \log\left(\frac{f_{\mu_0,\sigma}}{f_{\mu,\sigma}}\right)^2. \tag{6.8}$$

As mentioned before, we work with a compactly supported approximation $f_{0\sigma}$ of f_0 on E_σ, with the support E'_σ of $f_{0\sigma}$ containing E_σ and exploit the exponentially decaying tails of f_0 on E_σ^c. m_σ in 6.8 is an infinitely smooth function whose choice will be made explicit later. We now provide a detailed analysis of each term in 6.8.

We start with the last term on the right hand side of 6.8. The main idea is to work on a norm-bounded subset of the function space where the density function $f_{\mu,\sigma}$ can be bounded below and utilize the sub-Gaussian tails of $\|\mu\|_\infty$ to bound the integral outside the above region. Observe that for $\|\mu\|_\infty \le M$,

$$f_{\mu,\sigma}(y) \ge \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(|y|+M)^2},$$

which implies

$$\begin{aligned}
&\int_{E_\sigma^c} f_0(y) \log\left(\frac{f_{\mu_0,\sigma}(y)}{f_{\mu,\sigma}(y)}\right)^2 dy \le \int_{E_\sigma^c} f_0(y) \log\left\{\frac{C_4\sigma}{e^{-1/(2\sigma^2)(y-M)^2}}\right\} dy \\
\le\ & \log C_4\sigma \int_{E_\sigma^c} f_0(y)dy + \frac{1}{2\sigma^2}\int_{E_\sigma^c}(y-M)^2 dy \\
=\ & \left(\log C_4\sigma + \frac{M^2}{2\sigma^2}\right)\int_{E_\sigma^c} f_0(y)dy + \frac{1}{2\sigma^2}\int_{E_\sigma^c} y^2 f_0(y)dy - \frac{M}{\sigma^2}\int_{E_\sigma^c} y f_0(y)dy.
\end{aligned}$$

Now since $\int_{E_\sigma^c} y^j f_0(y)dy \le \sigma^{H_1/2}\int_{E_\sigma^c} y^j\sqrt{f_0(y)}dy, j = 0,1,2$, we need to choose H_1 satisfying $\sigma^{H_1/2} \approx \sigma^6/M$ to make $\int_{E_\sigma^c} f_0(y)\log\left(\frac{f_{\mu_0,\sigma}(y)}{f_{\mu,\sigma}(y)}\right)^2 dy \precsim O(\sigma^4)$.

To bound the integral over the set $\{\|\mu\|_\infty > M\}$, we provide an upper bound to $P(\|W^J\|_\infty > M)$ in the following Lemma 6.1, with proof in the Appendix.

Lemma 6.1. *With* $\sigma_J^2 = \sum_{j=1}^J \lambda_j^2$,

$$P(\|W^J\|_\infty > M) \le 2aM\exp\left[-\frac{1}{2b^2\sigma_J^2}\left\{M - C_6\frac{1}{M}\{(\log a)^{1/2} + (\log M)^{1/2}\}\right\}^2\right]$$

for some constant $C_6 > 0$.

We now consider the three terms inside the parenthesis in 6.8. Let us start with $\int_{E_\sigma} f_0 \log(f_{\mu_0,\sigma}/f_{\mu_{0\sigma},\sigma})^2$. For $y \in E_\sigma$, $f_{\mu_0,\sigma}(y)/f_{\mu_{0\sigma},\sigma}(y) = \phi_\sigma * f_0(y)/\phi_\sigma * f_{0\sigma}(y)$. Recall, $f_{0\sigma}(y) = f_0 1_{E'_\sigma}(y)/\psi_\sigma$ with $\psi_\sigma = \int_{E'_\sigma} f_0(y)$. Note that

$$\begin{aligned}\psi_\sigma \geq 1 - \int_{E'_\sigma} f_0(x)dx &\geq 1 - \sigma^{H_2/2} \int_{E'_\sigma} \sqrt{f_0(x)}dx \\ &\geq 1 - \sigma^4\end{aligned}$$

for $H_2 \geq 8$. Now,

$$\begin{aligned}&\phi_\sigma * f_0(y) = \int \phi_\sigma(y-t) f_0(t)dt \\ &= \int_{E'_\sigma} \phi_\sigma(y-t) f_0(t)dt + \int_{(E'_\sigma)^c} \phi_\sigma(y-t) f_0(t)dt.\end{aligned}$$

Hence,

$$\frac{\phi_\sigma * f_0(y)}{\phi_\sigma * f_{0\sigma}(y)} = \psi_\sigma \left\{ 1 + \frac{\int_{(E'_\sigma)^c} \phi_\sigma(y-t) f_0(t)dt}{\int_{E'_\sigma} \phi_\sigma(y-t) f_0(t)dt} \right\}.$$

Now, for $t \in (E'_\sigma)^c$ and $y \in E_\sigma$, $f_0(t) \leq \sigma^{H_2} \leq \sigma^{H_2-H_1} f_0(y)$, implying $\int_{(E'_\sigma)^c} \phi_\sigma(y-t) f_0(t)dt \leq \sigma^{H_2-H_1} f_0(y)$. Moreover,

$$\begin{aligned}&\int_{E'_\sigma} \phi_\sigma(y-t) f_0(t)dt = \phi_\sigma * f_0(y) - \int_{(E'_\sigma)^c} \phi_\sigma(y-t) f_0(t)dt \\ &\geq C f_0(y) - \sigma^{H_2-H_1} f_0(y).\end{aligned}$$

Thus,

$$\frac{\phi_\sigma * f_0(y)}{\phi_\sigma * f_{0\sigma}(y)} \precsim \psi_\sigma \left(1 + \frac{\sigma^{H_2-H_1}}{C - \sigma^{H_2-H_1}} \right) \precsim 1 \tag{6.9}$$

On the other hand,

$$\begin{aligned}&\frac{\phi_\sigma * f_0(y)}{\phi_\sigma * f_{0\sigma}(y)} = \psi_\sigma \frac{\phi_\sigma * f_0(y)}{\int_{E'_\sigma} \phi_\sigma(y-t) f_0(t)dt} \\ &\geq \psi_\sigma = 1 + O(\sigma^4).\end{aligned} \tag{6.10}$$

Hence, from 6.9 & 6.10, one has $\phi_\sigma * f_0(y)/\phi_\sigma * f_{0\sigma}(y) = 1 + O(\sigma^4)$ for $y \in E_\sigma$, implying $\int_{E_\sigma} f_0 \log(\phi_\sigma * f_0/\phi_\sigma * f_{0\sigma})^2 = O(\sigma^4)$.

We next turn our attention to $\int_{E_\sigma} f_0 \log(f_{\mu_{0\sigma},\sigma}/f_{m_\sigma,\sigma})^2$. Ghosal and van der Vaart (2007) showed that a Gaussian convolution of a compactly supported distribution can be approximated with high accuracy by a finite mixture of normals with "relatively few"

mixture components and Kruijer, Rousseau and van der Vaart (2010) obtained a finer calibration of their result to handle the above integral. However, it becomes unwieldy to use their result in our setup since their approximating density is obtained as the convolution of a Gaussian kernel with a discrete mixing distribution with finitely many support points, with the corresponding quantile function being a step function on $[0,1]$ with finitely many jumps. Although one can place a prior on μ whose realizations are step functions, several issues arise with the posterior computation including choosing & updating the number of steps. It would be appealing to use a smooth prior on μ and yet obtain a similar approximation result. We borrow techniques from the physics literature on maximum entropy moment matching or MAXENT (Mead and Papanicolaou, 1984) to develop an approximation result with a smooth mixing measure as follows:

Lemma 6.2. *Let f be a density compactly supported on $[-a_\sigma, a_\sigma]$ with $a_\sigma = a_0|\log\sigma|^{1/\tau_2}$ with σ small enough. Then, for any $A_0 > 0$, there exists an infinitely smooth density f_{m_σ} on $[-a_\sigma, a_\sigma]$, such that $\|\phi_\sigma * f - \phi_\sigma * f_{m_\sigma}\|_\infty \leq \sigma^{-1}\exp(-C|\log\sigma|^{1/\tau_2})$.*

A proof of Lemma 6.2 can be found in the Appendix.

The tail behavior of f_0 implied by Assumptions 3.2 & 6.1 imply $E'_\sigma \subset [-a_\sigma, a_\sigma]$ with $a_\sigma = a_0|\log\sigma|^{1/\tau_2}$ with $a_0 = \left(\frac{H_2}{2\tau_1}\right)^{1/\tau_2}$.

Let m_σ be the quantile function of the compactly supported density f_{m_σ} in Lemma 6.2 and let $f_{m_\sigma,\sigma} = \phi_\sigma * f_{m_\sigma}$. Note that for $y \in E'_\sigma$,

$$\begin{aligned} f_{\mu_{0\sigma},\sigma}(y) &= \frac{1}{\psi_\sigma}\int_{E'_\sigma}\phi_\sigma(y-t)f_0(t)dt \\ &\geq \frac{(C-\sigma^{H_2-H_1})}{1-\sigma^4}f_0(y) \geq \frac{C}{2}\sigma^{H_2}, \end{aligned}$$

for sufficiently small σ. Using $|\log x| \leq \max\{\log|x-1|, \log|1/x-1|\}$ for $x > 0$, one gets

$$\int_{E_\sigma} f_0 \log\left(\frac{f_{\mu_{0\sigma},\sigma}}{f_{m_\sigma,\sigma}}\right)^2 \leq \int_{E_\sigma} f_0\left(\frac{\|f_{\mu_{0\sigma},\sigma} - f_{m_\sigma,\sigma}\|_\infty}{(C/3)\sigma^{H_2}}\right)^2,$$

for A_0 large enough. By choosing A_0 sufficiently large and using Lemma 6.2, we obtain

$$\left(\frac{\|f_{\mu_{0\sigma},\sigma} - f_{m_\sigma,\sigma}\|_\infty}{(C/3)\sigma^{H_2}}\right)^2 \precsim O(\sigma^4). \tag{6.11}$$

Finally, we consider the third term $\int f_0(\log f_{m_\sigma,\sigma}/f_{\mu,\sigma})^2$ inside the parenthesis in 6.8. Proceeding as in the previous case, we first lower bound $f_{m_\sigma,\sigma}$ on E'_σ. In the previous case, we already obtained $f_{\mu_{0\sigma},\sigma} \succsim \sigma^{H_2}$. From Borwein and Lewis (1991), $\|f_{m_\sigma} - f_{0\sigma}\|_\infty = o(k^{-1})$ if we match k moments. From Lemma 6.2, we know that $k \approx O(\sigma^{-\alpha}|\log\sigma|^{\alpha/\tau_2})$ and hence by choosing α large enough, we can make $f_{m_\sigma,\sigma} \succsim \sigma^{H_2}$ on E'_σ. Now, using the same bound for $|\log x|$ as before, we need $\left\|f_{m_\sigma,\sigma} - f_{\mu,\sigma}\right\|_\infty$ to be

$O(\sigma^H)$ for any $H > 0$. From Kruijer, Rousseau and van der Vaart (2010) it follows that $\sup_y |\phi_\sigma(y-\mu_1) - \phi_\sigma(y-\mu_2)| \precsim \frac{|\mu_1-\mu_2|}{\sigma^2}$ so that

$$\left\|f_{m_{\sigma,\sigma}} - f_{\mu,\sigma}\right\|_\infty \precsim \frac{\|\mu - m_\sigma\|_\infty}{\sigma^2}. \tag{6.12}$$

We shall now exploit the infinite differentiability of m_σ to view it as an element of $W_{2,q}$ for some large q and calculate the probability of the Gaussian process W^J concentrating around m_σ.

The reproducing kernel Hilbert space (RKHS) of W^J consists of the set of functions $w(t) = \sum_{j=0}^J w_j b\phi_j(at), t \in [0,1]$ with RKHS norm

$$\|w\|_{\mathbb{H}}^2 = \sum_{j=0}^J \frac{w_j^2}{\lambda_j^2}.$$

Since m_σ is infinitely differentiable, $m_\sigma \in W_{2,q}^C[0,1]$ for any $q \geq 1$ (with C possibly depending on q). Hence there exists $\{m_{\sigma,j}\}_{j=0}^\infty$ such that

$$m_\sigma(t) = \sum_{j=0}^\infty m_{\sigma,j} b\phi_j(at),\ t \in [0,1].$$

Consider the projection of m_σ on the RKHS of W^J as

$$m_\sigma^J(t) = \sum_{j=0}^J m_{\sigma,j} b\phi_j(at).$$

In the sequel, we will choose a $q \geq 1$ and sequences of integers $J_n \uparrow \infty$, a_n, b_n to achieve the optimal rate of convergence. To that end, we calculate the prior concentration of W^J around m_σ^J for a fixed J with $\lambda_j = j^{-q/4}$ for $j \geq 1$. Recall that the prior concentration function of the Gaussian process W^J around m_σ^J is given by

$$\phi_{m_\sigma^J}(\epsilon) = \inf_{h \in \mathbb{H}: \|h - m_\sigma^J\|_\infty < \epsilon} \|h\|_{\mathbb{H}}^2 - \log \mathrm{P}(\left\|W^J\right\|_\infty \leq \epsilon). \tag{6.13}$$

Lemma 6.3. *For $q > 16$,*

$$\phi_{m_\sigma^J}(\epsilon) \precsim \frac{1}{a^q b^2} \left\|m_\sigma^J\right\|_{2,q}^2 + \begin{cases} J\left(1 + \log \frac{b}{\epsilon}\right),\ \epsilon J^{q/4} \precsim J^2 \\ \left(\frac{b}{\epsilon}\right)^{20/q},\ \epsilon J^{q/4} \geq J^2. \end{cases}$$

For a proof, refer to the Appendix.

Recall that we need the concentration bound for m_σ, while in the above lemma, we obtained the concentration bound for m_σ^J. We thus need error bounds on how well the

truncation m_σ^J approximates the function m_σ. Noting that $m_{j,\sigma}^2 \leq \frac{1}{b^2}\|m_\sigma\|_{2,q}^2 (aj)^{-q}$,

$$\begin{aligned}\left\|m_\sigma - m_\sigma^J\right\|_\infty &\leq \left\|\sum_{j=J+1}^{\infty} m_{j,\sigma}\phi_j(t)\right\|_\infty \\ &\leq b\sum_{j=J+1}^{\infty}|m_{j,\sigma}| \\ &\leq \|m_\sigma\|_{2,q}(aJ)^{-(q/2-1)}. \end{aligned} \tag{6.14}$$

To bound the final term in (6.14), we provide an upper bound to $\|m_\sigma\|_{2,q}^2$ in the following Lemma 6.4.

Lemma 6.4. $\|m_\sigma\|_{2,q}^2 \leq \sigma^{-(2q-1)H_2}$.

Proof. Exact derivation of the bound is quite tedious and we shall only sketch the main steps of the proof. Recall that $m_\sigma(x) = F_{m_\sigma}^{-1}(x), x \in [0,1]$ where $F_{m_\sigma} : [-a_\sigma, a_\sigma] \to [0,1]$ given by $F_{m_\sigma}(x) = \int_{-a_\sigma}^{x} f_{m_\sigma}(t)dt$. Then

$$\|m_\sigma\|_{2,q}^2 = \int_0^1 \{(F_{m_\sigma}^{-1})^{(q)}(x)\}^2 dx \tag{6.15}$$

Observe that

$$\begin{aligned}(F_{m_\sigma}^{-1})'(1) &= \frac{1}{f_{m_\sigma}(a_\sigma)}, (F_{m_\sigma}^{-1})''(1) = -\frac{f'_{m_\sigma}(a_\sigma)}{f_{m_\sigma}(a_\sigma)^3} \\ (F_{m_\sigma}^{-1})'''(1) &= -3\frac{f'_{m_\sigma}(a_\sigma)^2}{f_{m_\sigma}(a_\sigma)^5} - \frac{f''_{m_\sigma}(a_\sigma)}{f_{m_\sigma}(a_\sigma)^4}.\end{aligned}$$

Proceeding like this, one has $\{f_{m_\sigma}(a_\sigma)\}^{2q-1}$ in the denominator for $(F_{m_\sigma}^{-1})^{(q)}(1)$ and $\{f_{m_\sigma}(-a_\sigma)\}^{2q-1}$ for $(F_{m_\sigma}^{-1})^{(q)}(0)$. The numerator terms of the above expression are bounded. From Borwein and Lewis (1991), we know that $\|f_{m_\sigma} - f_{0\sigma}\|_\infty = o(k^{-1})$ if we match k moments. From Lemma 6.2, $k \approx O(\sigma^{-\alpha}|\log\sigma|^{\alpha/\tau_2})$ and hence by choosing α large enough we can make $f_{m_\sigma}(a_\sigma) \geq f_{0\sigma}(a_\sigma) - \sigma^{H_2+1}$ and $f_{m_\sigma}(-a_\sigma) \geq f_{0\sigma}(-a_\sigma) - \sigma^{H_2+1}$ which implies,

$$(F_{m_\sigma}^{-1})^{(q)}(1) \precsim \sigma^{-(2q-1)H_2}, (F_{m_\sigma}^{-1})^{(q)}(0) \precsim \sigma^{-(2q-1)H_2}. \tag{6.16}$$

Noting that $\int_0^1 \{(F_{m_\sigma}^{-1})^{(q)}(x)\}^2 dx \precsim \max\{\{(F_{m_\sigma}^{-1})^{(q)}(0)\}^2, \{(F_{m_\sigma}^{-1})^{(q)}(1)\}^2\}$, the conclusion follows immediately. □

We are now in a position to complete the proof of Theorem 6.1.

6.1. Proof of Theorem 6.1

Proof. Since we can bound the numerator of the rhs of 6.12 as

$$\left\|W^J - m_\sigma\right\|_\infty \leq \left\|W^J - m_\sigma^J\right\|_\infty + \left\|m_\sigma - m_\sigma^J\right\|_\infty, \tag{6.17}$$

we need $\left\|m_\sigma - m_\sigma^J\right\|_\infty = \|m_\sigma\|_{2,q} (aJ)^{-(q/2-1)}$ to be $O(\sigma^{H_2+4})$ so that the fourth term of 6.8 is $O(\sigma^4)$. Next, we calculate the prior probability $P(\left\|W^J - m_\sigma^J\right\|_\infty \leq \sigma^{H_2+4})$. Using Lemma 6.3, we can see that if

$$M_n \sigma_n^{H_1/2} = O(\sigma_n^6), \tag{6.18}$$

$$P(\left\|W^J\right\|_\infty > M_n) = O(e^{-1/\sigma_n}), \tag{6.19}$$

$$\|m_\sigma\|_{2,q} (a_n J_n)^{-(q/2-1)} = O(\sigma_n^{H_2+4}), \tag{6.20}$$

$$(b_n \sigma_n^{-(H_2+4)})^{20/q} = O(\sigma_n^{-1}), \tag{6.21}$$

$$\frac{1}{a_n^q b_n^2} \left\|m_\sigma^J\right\|_{2,q}^2 = O(\sigma_n^{-1}), \tag{6.22}$$

$$J_n = O(\sigma_n^{-1}), \tag{6.23}$$

and $\sigma \in [\sigma_n, 2\sigma_n]$, then

$$\begin{aligned} &\Pi\left(f_{\mu,\sigma} : \int f_0 \log \frac{f_0}{f_{\mu,\sigma}} \leq \sigma_n^4, \int f_0 \log \left(\frac{f_0}{f_{\mu,\sigma}}\right)^2 \leq \sigma_n^4\right) \\ &\geq P(\sigma \in [\sigma_n, 2\sigma_n], \left\|W^J - m_\sigma^J\right\|_\infty \leq \sigma_n^{H_2+4}, \left\|W^J\right\|_\infty \leq M_n) \\ &\geq O(\sigma_n^{-1} |\log \sigma_n|). \end{aligned}$$

Next we make specific choices for q, a_n, b_n, M_n and σ_n. Clearly $\sigma_n = O(n^{-1/5})$ for the optimal rate. 6.18-6.23 determine the values of the sequences M_n, a_n, b_n and q using the upper bounds on $\|m_\sigma\|_{2,q}$ and $P(\left\|W^J\right\|_\infty > M_n)$ provided in Lemma 6.4 and Lemma 6.1 respectively. It can be verified that 6.20, 6.21 and 6.22 are satisfied with q greater than the positive root of the quadratic equation

$$(9/10)q^2 - (10H_2 + 24/5)q - 2(2H_2 + 7) = 0, \tag{6.24}$$

which is satisfied by $q \approx 95$. Choosing $q = 150$, $H_1 = H_2 = 12$, $a_n = n^\alpha$ for some $\alpha > 1$, $M_n^2 = O(\log n), b_n = O(n^{-1/10} (\log n)^{1/2})$, we can see that the convergence rate is $\epsilon_n = n^{-2/5} \log^{t_0} n$ for some global constant t_0. □

7. Special cases

A desirable property of any nonparametric model is that it can "collapse" back to a simpler structure when the additional flexibility is not warranted. For example, the nonparametric prior may be centered on a smaller class of densities (e.g., a parametric family), with a faster rate of convergence obtained when the true density falls within this smaller class. In this section, we study such collapsing behavior in a couple of cases.

7.1. Properly centering the prior leads to parametric rate of convergence

We have already noted in the introduction that one can center the non-parametric model on a parametric family by centering the prior on the transfer function on a parametric class of quantile (or inverse c.d.f.) functions $\{F_\theta^{-1} : \theta \in \Theta\}$. Here we show that if our guess about the true density f_0 is correct, we can actually achieve a parametric rate of convergence by centering the prior for μ on F_0^{-1}. Centering the prior on the true quantile function expands the RKHS to include the best approximation which is the true quantile function itself. We formalize our result in the following Theorem 7.1.

Assumption 7.1. *Define μ_0 to be F_0^{-1}. We assume μ follows a Gaussian process $GP(\mu_0, c)$ centered at μ_0 and with a squared exponential covariance kernel $c(\cdot,\cdot;A)$ and a Gamma prior for the inverse-bandwidth A, so that, $c(t,s;A) = e^{-A(t-s)^2}, t,s \in [0,1], A \sim Ga(p,q)$.*

Assumption 7.2. $\frac{\sigma^2}{n^{-1/4}\log n} \sim IG(a,b)$.

Theorem 7.1. *If f_0 is compact and satisfies Assumption 3.1 and the priors Π_μ and Π_σ are as in Assumptions 7.1 and Assumption 7.2 respectively with the correct centering f_0, the best obtainable posterior rate of convergence relative to h is*

$$\epsilon_n = n^{-\frac{1}{2}}(\log n)^{t_0}. \tag{7.1}$$

for some global constant t_0.

Proof. The portion from which the proof differs from the proof of Theorem 5.1 is the calculation of the prior concentration. Let $\mu = \mu_0 + W$ where $W \sim \mathrm{GP}(0,c)$. It is easy to see that

$$P(\|\mu - \mu_0\|_\infty \le 2\epsilon) = P(\|W\|_\infty \le 2\epsilon) \ge e^{-C_1 a_1 \log^2(a_1/\epsilon)} P(a_0 < A < a_1).$$

Hence with $\sigma_n = n^{-1/4}, l_n = n^{-1/4}, h_n = n^\beta$ for some $\beta > 0$, we can show that

$$P(\sigma \in [\sigma_n, 2\sigma_n]) \ge \exp\{-(\log n)^{t_1}\}$$

for some $t_1 > 0$.

Define the Gaussian process sieve to be

$$B_n = f_0 + \left[\left(M_n\sqrt{\frac{r_n}{\xi_n}}\mathbb{H}_1^r + \bar{\delta}_n\mathbb{B}_1\right) \cup \left(\cup_{a<\xi_n}\left(M_n\mathbb{H}_1^a\right) + \bar{\delta}_n\mathbb{B}_1\right)\right],$$

where the sequences $\xi_n, \bar{\delta}_n, M_n$ are exactly as specified in the proof of Theorem 5.1. It follows from the proof of Theorem 5.1 that $\epsilon_n = n^{-1/2}(\log n)^{t_0}$ for some constant $t_0 > 0$. □

***Remark* 7.1.** The extension to the case when the true density is actually non-compact and satisfies the tail condition in Assumption 6.1 can be handled following the steps in the proof of Theorem 6.1 with Π_μ in Assumption 6.2 centered at the non-compact f_0 for appropriate choices of sequences a_n, b_n and λ_j and Π_σ as in Assumption 5.2. However one can show that the best obtainable rate of convergence can only be made arbitrarily close to the parametric rate in the sense that the rate of convergence would be slower compared to the parametric rate by a factor n^β where β can be arbitrarily small.

***Remark* 7.2.** A more interesting and practical extension of Theorem 7.1 is the case where one has correctly guessed a parametric family which contains the true density. Suppose the parametric family is given by $\{f_\theta : \theta \in \mathbb{R}^p\}$ indexed by the parameter θ living in some Euclidean space, and the true density $f_0 = f_{\theta_0}$. It is natural then to center the prior for μ on F_θ^{-1}, with a hyperprior on θ quantifying uncertainty about the value of the finite-dimensional parameter θ. A straightforward application of Theorem 7.1 shows that it is possible to attain parametric rate of convergence under the same assumptions as in Theorem 7.1 if the prior on θ has full support on $\mathbb{R}^p$ and the mapping $\theta \to f_\theta$ satisfies mild regularity conditions, e.g., as in Ghosal, Ghosh and van der Vaart (2000). In particular, one obtains the near parametric rate if $\theta \sim N_p(\mu_0, \Sigma_0)$. If the prior guess about the parametric family is incorrect, then one would still get the near minimax rate for the class of twice continuously differentiable densities.

7.2. True density is a Gaussian convolution with a finite mixture of truncated Gaussians

Ghosal and van der Vaart (2001) showed that when the true density is a location-scale or location mixture of normals $\int \frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right) F_0(\mu, \sigma)$ with the scale parameter lying between two fixed numbers and the mixing distribution F_0 being either compactly supported or having sub-Gaussian tails, a Dirichlet process mixture of normals can achieve near-parametric rate of convergence. To mimic the above super-smooth case for our non-linear latent variable model, we shall consider a simplistic situation when the true density is a Gaussian convolution with a finite mixture of truncated Gaussians with the same truncation bounds. We show below that the rate of convergence in that case can be as close as possible to the parametric rate. The actual super-smooth case would be the situation when the true density is a finite mixture of Gaussians, so that it can be expressed as a convolution with a finite mixture of Gaussians. Remark 7.3 discusses very briefly about that case.

Theorem 7.2. *Given any $\alpha > 0$. If f_0 is $\phi_{\sigma_0} * f_1$ where f_1 is a finite mixture of truncated Gaussians with the same truncation bounds and the prior Π_μ is as in Assumptions 5.1 and $\frac{\sigma}{n^{-1/(2\alpha+1)}\log n} \sim IG(a, b)$ respectively, then the best obtainable rate of posterior convergence is*

$$n^{-\frac{\alpha}{2\alpha+1}} \log^{t_\alpha}(n) \tag{7.2}$$

where t_α is a constant depending on α.

Proof. Clearly f_1 is an infinitely smooth density which has quantile function $\mu_1 = F_1^{-1}$ infinitely smooth in $[0,1]$ and hence $\mu_1 \in C^\alpha([0,1])$ for any $\alpha > 0$. Observe that

$$\begin{aligned} h^2(f_0, f_{\mu,\sigma}) &= h^2(\phi_{\sigma_0} * f_1, f_{\mu,\sigma}) \\ &= h^2(f_{\mu_1,\sigma_0}, f_{\mu,\sigma_0}) + h^2(f_{\mu,\sigma_0}, f_{\mu,\sigma}) \\ &\precsim \|\mu - \mu_1\|_\infty^2 + \frac{|\sigma_0 - \sigma|}{\sigma} \end{aligned}$$

From van der Vaart and van Zanten (2009), we obtain $P(\|\mu - \mu_1\|_\infty < \sigma_n^\alpha) \geq \exp(-\sigma_n^{-1})$ and $P(|\sigma_0 - \sigma| < \sigma_n^{2\alpha+1}) \geq \exp(-\sigma_n^{-1})$ for $\sigma_n = n^{-1/(2\alpha+1)}$. With the same sieve as in the proof of Theorem 5.1, it follows that $\epsilon_n = n^{-\alpha/(2\alpha+1)}(\log n)^{t_\alpha}$ for some constant $t_\alpha > 0$. □

***Remark* 7.3.** The extension to the case when the true density is a finite mixture of Gaussians can be handled following the steps in the proof of Theorem 6.1 with Π_μ and Π_σ in Assumptions 6.2 and Assumption 5.2 respectively for appropriate choices of sequences a_n, b_n and λ_j respectively.

8. Discussion

Non-linear latent variable models offer a flexible modeling framework in a broad variety of problems and improved practical performance has been demonstrated by Lawrence (2004, 2005); Lawrence and Moore (2007); Ferris, Fox and Lawrence (2007); Kundu and Dunson (2011) among others. The univariate density estimation model studied here can be extended to multivariate density estimation, latent factor modeling and density regression problems; we are currently studying theoretical properties of these extensions building upon the results developed in this article in the baseline case.

In standard Gaussian process regression, the regression function is assumed to be continuous on a compact domain, and one can use standard results on concentration bounds for Gaussian processes (van der Vaart and van Zanten, 2008b). However, we cannot use these results directly as the quantile function of a density supported on the entire real line is unbounded near zero and one. To address this problem, we required assumptions on the tails of the true density and exploited the interplay between the tails of a density and the boundary behavior of the corresponding quantile function. Building a sequence of compact approximations to the true density, accurate concentration bounds around the corresponding quantile functions (which are in $C[0,1]$) are developed for the Gaussian process prior on the transfer function. While deriving this bound, one has to carefully calibrate the rate at which the RKHS norms of the sequence of approximating quantile functions increase to infinity. A truncated series prior is convenient for this purpose, however one needs to appropriately rescale the prior as in 6.2 for optimal rate. It would be interesting to study whether one obtains the same for a host of other commonly used

Gaussian process priors. It should be noted here that posterior consistency with commonly used Gaussian process priors is immediate using our treatment of the non-compact case.

We finally note that although our results assume twice continuously differentiability of the true density, one can obtain optimal rate of convergence for arbitrary degree of smoothness of the truth. From the treatment of $\int f_0 \log(f_0/f_{\mu,\sigma})^2$ in (6.8), it is not difficult to see that all the terms barring $\int f_0 \log(f_0/f_{\mu_0,\sigma})^2$ can be made $O(\sigma^H)$ for arbitrarily large H. The first term $\int f_0 \log(f_0/f_{\mu_0,\sigma})^2$ cannot be improved beyond $O(\sigma^4)$ even if the true density is more than twice continuously differentiable, since $h(f_0, \phi_\sigma * f_0)$ can only be $O(\sigma^2)$. This is a well-known issue with Gaussian convolutions and one can improve the approximation bound by using a higher order kernel ψ_σ (Fan and Hu, 1992; Marron and Wand, 1992), so that $\|f_0 - \psi_\sigma * f_0\| = O(\sigma^H)$ for H arbitrarily large. A thorny issue with using higher-order kernels in the frequentist literature is that $\psi_\sigma * f_0$ is not guaranteed to be positive everywhere, but one can bypass that easily in a Bayesian framework as one only needs to show that the prior support contains densities that are appropriately close to the true density. Letting $\phi_\sigma * f_1 = \psi_\sigma * f_0$, one can solve for f_1 using inverse-Fourier transforms and one has $\|f_0 - \phi_\sigma * f_1\| = O(\sigma^H)$. Although $\int f_1 = 1$, f_1 can be negative at some places. Kruijer, Rousseau and van der Vaart (2010) showed that under suitable conditions on the true density, one obtains the same approximation error for f_2, the positive part of f_1 normalized to integrate to one. Kruijer, Rousseau and van der Vaart (2010) used the twicing kernel method (Newey, Hsieh and Robins, 2004) to obtain f_1 in a closed-form, one can use the same trick here or use other higher-order kernels to obtain f_1. One can then simply replace μ_0 with the quantile function of f_2 and proceed with the rest of the analysis identically.

Appendix

A.1. Proof of Theorem 4.1

Proof. Let f_0 be a density with quantile function μ_0 that satisfies the conditions of Theorem 4.1. Observe that $\|\mu_0\|_1 = \int_{t=0}^1 |\mu_0(t)|\, dt = \int_{-\infty}^{\infty} |z|\, f_0(z) dz < \infty$ since f_0 has a finite first moment, and thus $\mu_0 \in \mathrm{L}_1[0,1]$. Fix $\epsilon > 0$. We want to show that $\Pi\{B_\epsilon(f_0)\} > 0$, where $B_\epsilon(f_0) = \{f \ : \ \|f - f_0\|_1 < \epsilon\}$.

Note that $\mu_0 \notin C[0,1]$, so that $\mathrm{pr}(\|\mu - \mu_0\|_\infty < \epsilon)$ can be zero for small enough ϵ. The main idea is to find a continuous function $\tilde{\mu_0}$ close to μ_0 in L_1 norm and exploit the fact that the prior on μ places positive mass to arbitrary sup-norm neighborhoods of $\tilde{\mu_0}$. The details are provided below.

Since $\|\phi_\sigma * f_0 - f_0\|_1 \to 0$ as $\sigma \to 0$, find σ_1 such that $\|\phi_\sigma * f_0 - f_0\|_1 < \epsilon/2$ for $\sigma < \sigma_1$. Pick any $\sigma_0 < \sigma_1$. Since $C[0,1]$ is dense in $\mathrm{L}_1[0,1]$, for any $\delta > 0$, we can find a continuous function $\tilde{\mu_0}$ such that $\|\mu_0 - \tilde{\mu_0}\|_1 < \delta$. Now, $\|f_{\mu,\sigma} - f_{\tilde{\mu_0},\sigma}\|_1 \leq C \|\mu - \tilde{\mu_0}\|_1 / \sigma$

for a global constant C. Thus, for $\delta = \epsilon\sigma_0/4$,

$$\big\{f_{\mu,\sigma} \ : \ \sigma_0 < \sigma < \sigma_1, \|\mu - \tilde{\mu_0}\|_\infty < \delta\big\} \subset \big\{f_{\mu,\sigma} \ : \ \|f_0 - f_{\mu,\sigma}\|_1 < \epsilon\big\},$$

since $\|f_0 - f_{\mu,\sigma}\|_1 < \|f_0 - f_{\mu_0,\sigma}\|_1 + \|f_{\mu_0,\sigma} - f_{\tilde{\mu_0},\sigma}\|_1 + \|f_{\tilde{\mu_0},\sigma} - f_{\mu,\sigma}\|_1$ and $f_{\mu_0,\sigma} = \phi_\sigma * f_0$. Thus, $\Pi\{B_\epsilon(f_0)\} > \mathrm{pr}(\|\mu - \tilde{\mu_0}\|_\infty < \delta)\ \mathrm{pr}(\sigma_0 < \sigma < \sigma_1) > 0$, since Π_μ has full sup-norm support and Π_σ has full support on $[0, \infty)$. □

A.2. Proof of Lemma 6.1

Proof. From Theorem 5.2 of Adler (1990) it follows that if X is a centered Gaussian process on a compact set $T \subset \mathbb{R}^d$ and σ_T^2 is the maximum variance attained by the Gaussian process on T, then for large M,

$$P(\|X\|_\infty > M) \leq 2N(1/M, T, \|\cdot\|) \exp\left[-\frac{1}{2\sigma_T^2}\{M - \nu(M)\}^2\right], \tag{A.1}$$

where $\nu(M) = C_5 \int_0^{1/M} \{\log N(1/M, T, \|\cdot\|)\}^{1/2} d(1/M)$ for some constant $C_5 > 0$. Observe that W^J is rescaled to $T = [0, a]$ and the maximum variance attained by W^J is $b^2\sigma_J^2$. Note that $N(1/M, T, \|\cdot\|) = aM$. Now

$$\begin{aligned}
\nu(M) &\leq C_6 \int_0^{1/M} \{\log(aM)\}^{1/2} d(1/M) \\
&\leq C_6 \int_0^{1/M} \{(\log a)^{1/2} + (\log M)^{1/2}\} d(1/M) \\
&\leq C_6 \frac{1}{M}\{(\log a)^{1/2} + (\log M)^{1/2}\}
\end{aligned}$$

for some constant $C_6 > 0$. Plugging in the value of $N(1/M, T, \|\cdot\|)$ and the bound for $\nu(M)$ in A.1, we get the required bound for $P(\|W^J\|_\infty > M)$. □

A.3. Proof of Lemma 6.2

Proof. From Mead and Papanicolaou (1984), for any $k \geq 1$, we can get an infinitely smooth density f_{m_σ} supported on $[-a_\sigma, a_\sigma]$ such that

$$\int_{-a_\sigma}^{a_\sigma} x^j f_{m_\sigma}(x) dx = \int_{-a_\sigma}^{a_\sigma} x^j f(x) dx, \ j = 0, \ldots, k. \tag{A.2}$$

One possible choice of f_{m_σ} in A.2 has the form $f_{m_\sigma}(x) = \exp(-\sum_{l=1}^k b_l x^l)$ which corresponds to the maximum entropy moment matching (MAXENT) density. We shall choose k sufficiently large depending on σ so that one has the desired approximation result.

Consider an interval $I_\sigma = [-(a_\sigma + t_\sigma), (a_\sigma + t_\sigma)]$ containing the interval $[-a_\sigma, a_\sigma]$ for some $t_\sigma > 0$ to be chosen later depending on σ. Observe that

$$\sup_{x \in I_\sigma^c} |\phi_\sigma * f(x) - \phi_\sigma * f_{m_\sigma}(x)| \leq \sup_{x \in I_\sigma^c} \int_{-a_\sigma}^{a_\sigma} \phi_\sigma(x-y)\, |f(y) - f_{m_\sigma}(y)|\, dy \leq 2\phi_\sigma(t_\sigma). \quad \text{(A.3)}$$

Next, along the lines of Ghosal and van der Vaart (2007)

$$\begin{aligned} \sup_{x \in I_\sigma} \left| \int_{-a_\sigma}^{a_\sigma} \phi_\sigma(x-y) f(y) - f_{m_\sigma}(y) dy \right| &\leq 2 \sup_{x \in I_\sigma, |y| \leq a_\sigma} \left| \phi_\sigma(x-y) - \sum_{j=0}^{k} \frac{1}{\sqrt{2\pi}} \frac{(-1)^j \sigma^{-(2j+1)} (x-y)^{2j}}{j!} \right| \\ &\leq \frac{2C_1}{\sigma} \sup_{x \in I_\sigma, |y| \leq a_\sigma} \left(\frac{e|x-y|^2}{2k\sigma^2} \right)^k \\ &\leq C_2 \sigma^{-(2k+1)} \left(\frac{e}{2} \right)^k \frac{(2a_\sigma + t_\sigma)^{2k}}{k^k}, \end{aligned} \quad \text{(A.4)}$$

for some global constants $C_1, C_2 > 0$.

Now choose $t_\sigma = Aa_\sigma$ for some constant $A > 0$. Then, from A.3 and A.4, we obtain,

$$\|\phi_\sigma * f - \phi_\sigma * f_{m_\sigma}\|_\infty \leq \max \left[2\phi_\sigma(t_\sigma), \frac{C_2}{\sigma} \exp \left\{ 2k \log \frac{(2+A)a_0 |\log \sigma|^{1/\tau_2}}{\sigma} - k \log \frac{k}{B} \right\} \right],$$

where $B = e/2$. Choosing $k = B\sigma^{-\alpha} |\log \sigma|^{\alpha/\tau_2}$,

$$k \log \frac{k}{B} = B\sigma^{-\alpha} |\log \sigma|^{\alpha/\tau_2} \{\alpha \log(1/\sigma) + (\alpha/\tau_2) \log(|\log \sigma|)\},$$

and

$$2k \log \frac{(2+A)a_0 |\log \sigma|^{1/\tau_2}}{\sigma} = 2B\sigma^{-\alpha} |\log \sigma|^{\alpha/\tau_2} \{\log\{(2+A)a_0\} + (1/\tau_2) \log(|\log \sigma|)\}.$$

Clearly, if $\alpha \geq 2$ and σ is sufficiently small, $2k \log \frac{(2+A)a_0 |\log \sigma|^{1/\tau_2}}{\sigma} < k \log \frac{k}{B}$. Then by choosing $\alpha > 2$, we can make $2\phi_\sigma(t_\sigma) > \frac{C_2}{\sigma} \exp \left\{ 2k \log \frac{(2+A)a_0 |\log \sigma|^{1/\tau_2}}{\sigma} - k \log \frac{k}{B} \right\}$ and hence

$$\|\phi_\sigma * f - \phi_\sigma * f_{m_\sigma}\|_\infty \leq 2\phi_\sigma(t_\sigma) = \frac{C_3}{\sigma} \exp\{-(a_0 A)^2/2 |\log \sigma|^{2/\tau_2}\}. \quad \text{(A.5)}$$

Since A is arbitrary, the conclusion of the theorem follows.

□

A.4. Proof of Lemma 6.3

Proof. m_σ^J is contained in the RKHS of W^J,

$$\inf\{\|w\|_{\mathbb{H}}^2 : \|w - m_\sigma^J\|_\infty < \epsilon\} = \|m_\sigma^J\|_{\mathbb{H}}^2 = \sum_{j=0}^{J} \frac{m_{\sigma,j}^2}{\lambda_j^2}$$

$$= \sum_{j=0}^{J} j^{q/2} m_{\sigma,j}^2 \leq \sum_{j=0}^{J} j^q m_{\sigma,j}^2 \precsim \frac{a^q}{b^2} \|m_\sigma^J\|_{2,q}^2 .$$

Next we calculate $\mathrm{P}(\|W^J\|_\infty \leq \epsilon)$ using a technique similar to the proof on Theorem 4.5 in van der Vaart and van Zanten (2008a). For any numbers $\alpha_j \geq 0$ with $\sum_{j=0}^{J} \alpha_j \leq 1$,we have

$$\begin{aligned} \mathrm{P}(\|W^J\|_\infty \leq \epsilon) &\geq \mathrm{P}(\sum_{j=0}^{J} |Z_j \lambda_j b| < \epsilon) \\ &\geq \prod_{j=0}^{J} \mathrm{P}(|Z_j \lambda_j| < \alpha_j \epsilon / b). \end{aligned}$$

Now, define a function $f : [0, \infty) \to \mathbb{R}$ given by $f(y) = -\log \mathrm{P}(|Z| < y) = -\log\{2\Phi(y) - 1\}$, where $Z \sim \mathrm{N}(0, 1)$. f is a decreasing function and following van der Vaart and van Zanten (2008a), f is bounded above by a multiple of $1 + |\log y|$ for $y \in [0, c]$ and bounded above by a multiple of $e^{-y^2/2}$ for $y \geq c$ for some $c > 0$. Thus, with $\alpha_j = (K + j^2)^{-1}$ for a large constant $K > 0$,

$$\begin{aligned} -\log P(\|W^J\|_\infty \leq \epsilon) &\leq \sum_{j=1}^{J} f(\alpha_j \epsilon j^{q/4}/b) + f(\epsilon/(Kb)) \\ &\leq \int_1^J f\left(\frac{\epsilon x^{q/4}}{b(K + x^2)}\right) dx + f(\epsilon/b(K+1)) + f(\epsilon/(Kb)), \end{aligned}$$

where the last inequality in the above display follows from the fact that f is decreasing and the map $x \mapsto x^{q/4}/(K + x^2)$ is non-decreasing on $[1, \infty)$ for any $K > 0$ as long as $q > 4$. For ϵ small enough so that $\epsilon/(Kb) < c$, $f(\epsilon/(Kb)) < 1 + \log(Kb/\epsilon)$.

Now consider two cases to bound the integral in the last display. If $\epsilon J^{q/4} \leq (K + J^2)$, $\epsilon x^{q/4}/(K + x^2) \leq 1$ for $x \in [1, J]$. Hence in that case,

$$\int_1^J f\left(\frac{\epsilon x^{q/4}}{b(K + x^2)}\right) dx \leq \int_1^J \left(1 + \left|\log\left(\frac{\epsilon x^{q/4}}{b(K + x^2)}\right)\right|\right) dx \tag{A.6}$$

$$\leq \int_1^J \left(1 + \log \frac{b(K+1)}{\epsilon}\right) dx \leq J\left(1 + \log \frac{b(K+1)}{\epsilon}\right). \tag{A.7}$$

On the other hand, if $\epsilon J^{q/4}/b > (K + J^2)$,

$$\int_1^J f\left(\frac{\epsilon x^{q/4}}{K+x^2}\right)dx = \left(\frac{b}{\epsilon}\right)^{4/q} \times (4/q) \times \int_{\epsilon/b}^{\epsilon J^{q/4}/b} f\left(\frac{y}{K+(by/\epsilon)^{8/q}}\right)y^{4/q-1}dy. \quad \text{(A.8)}$$

The integral above is bounded by

$$\int_0^{b/\epsilon} f\left(\frac{y}{K+(b/\epsilon)^{16/q}}\right)y^{4/q-1}dy + \int_{b/\epsilon}^{\infty} f\left(\frac{y}{K+y^{16/q}}\right)y^{4/q-1}dy \quad \text{(A.9)}$$

$$\leq z^{4/q}\int_0^{b/(z\epsilon)} f(x)x^{4/q-1}dx + \int_0^{\infty} f\left(\frac{y}{K+y^{16/q}}\right)y^{4/q-1}dy \quad \text{(A.10)}$$

for $z = (K + (b/\epsilon)^{16/q})$. The first integral is bounded as $\epsilon \downarrow 0$ and the second integral is finite for $q > 16$. Hence for $q > 16$,

$$\phi_{m_\sigma^J}(\epsilon) \precsim \frac{1}{a^q b^2}\left\|m_\sigma^J\right\|_{2,q}^2 + \begin{cases} J\left(1+\log\frac{b}{\epsilon}\right), \ \epsilon J^{q/4} \precsim bJ^2 \\ \left(\frac{b}{\epsilon}\right)^{20/q}, \ \epsilon J^{q/4} \geq bJ^2 \end{cases}$$

□

References

Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for general Gaussian processes* **12**. Institute of Mathematical Statistics.

Borwein, J. M. and Lewis, A. S. (1991). Convergence of best entropy estimates. *SIAM Journal on Optimization* **1** 191.

Fan, J. and Hu, T. C. (1992). Bias correction and higher order kernel functions. *Statistics & probability letters* **13** 235–243.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1** 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2** 615–629.

Ferris, B., Fox, D. and Lawrence, N. (2007). WiFi-SLAM using Gaussian process latent variable models. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence* 2480–2485.

Ghosal, S., Ghosh, J. and Ramamoorthi, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics* **27** 143–158.

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28** 500–531.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *The Annals of Statistics* **29** 1233–1263.

Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics* **35** 697–723.

Kruijer, W., Rousseau, J. and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics* **4** 1225–1257.

Kundu, S. and Dunson, D. B. (2011). Single Factor Transformation Priors for Density Regression. *DSS Discussion Series*.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics* **20** 1222–1235.

Lawrence, N. D. (2004). Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in neural information processing systems 16: proceedings of the 2003 conference* **16** 329. The MIT Press.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *The Journal of Machine Learning Research* **6** 1783–1816.

Lawrence, N. D. and Moore, A. J. (2007). Hierarchical Gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning* 481–488. ACM.

Lenk, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association* **83** 509–516.

Lenk, P. J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika* **78** 531.

Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *The Annals of Statistics* 712–736.

Mauldin, R. D., Sudderth, W. D. and Williams, S. (1992). Polya trees and random distributions. *The Annals of Statistics* **20** 1203–1221.

Mead, L. R. and Papanicolaou, N. (1984). Maximum entropy in the problem of moments. *Journal of Mathematical Physics* **25** 2404–2417.

Newey, W. K., Hsieh, F. and Robins, J. M. (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica* **72** 947–962.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4** 639–650.

Tokdar, S. T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics* **67** 90–110.

Tokdar, S. T. (2007). Towards a faster implementation of density estimation with logistic Gaussian process priors. *Journal of Computational and Graphical Statistics* **16** 633–655.

van der Vaart, A. W. and van Zanten, J. H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electronic Journal of Statistics* **1** 433–448.

van der Vaart, A. and van Zanten, J. (2008a). Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics* **36** 1435–1463.

van der Vaart, A. and van Zanten, J. (2008b). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections* **3** 200–222.

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estima-

tion using a Gaussian random field with inverse Gamma bandwidth. *The Annals of Statistics* **37** 2655–2675.

WALKER, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation* **36** 45–54.